LEVEL II

14 ARI-RM-

Return

9 Research Memorandum 77-31

6

# RELATIVE MERITS OF SEVERAL MISSING DATA ESTIMATORS IN PERSONNEL SELECTION PROCEDURES

10 James L. Raney, Paul J. Duffy and Arthur C. F. Gilbert

11 Feb 78

12 13

16 2Q763731A768

AD A077948

PERSONNEL ACCESSION AND UTILIZATION TECHNICAL AREA





U. S. Army

Research Institute for the Behavioral and Social Sciences

February 1978

408 010



79 22 5 146

Best
Available
Copy

Army Project Number
2Q763731A768

Officer and NCO
Training & Utilization

**Research Memorandum 77-31**

RELATIVE MERITS OF SEVERAL MISSING DATA ESTIMATORS
IN PERSONNEL SELECTION PROCEDURES

James L. Raney, Paul J. Duffy, and Arthur C. F. Gilbert

William H. Helme, Supervisory Project Director

Submitted by:
Ralph R. Canter, Chief
Personnel Accession and Utilization Technical Area

February 1978



Approved by:

E. Ralph Dusek, Director
Individual Training and Performance
Research Laboratory

J. E. Uhlaner, Technical Director
U.S. Army Research Institute for
the Behavioral and Social Sciences

Research Memorandums are informal reports on technical research problems. Limited distribution is made, primarily to personnel engaged in research for the Army Research Institute.

# RELATIVE MERITS OF SEVERAL MISSING DATA ESTIMATORS IN PERSONNEL SELECTION PROCEDURES

## INTRODUCTION

The primary criterion for selection of recipients of Army ROTC scholarships is an applicant's rank on an order-of-merit list based on his Whole Man Score (WMS). The WMS is a linear combination of variables believed useful as predictors of criteria such as academic success, aptitude for military science, leadership potential, and predisposition toward a military career. Four predictors are used: (1) high school class standing score; (2) Scholastic Aptitude Test (SAT) or American College Test (ACT) score; (3) extracurricular, athletic, and leadership activity score; and (4) personal interview score. The predictors are assigned arbitrary weights of 3:3:3:1, respectively (TRADOC Circular 145-5, 1975).

This paper documents the psychometric rationale underlying the method of computing the high school class standing portion of the WMS. The method involves adjusting for class size by converting class rank to percentile rank. Percentile rank is then transformed to a normalized standard score by introducing an a priori distributional assumption. Several alternative methods for estimating missing class standing scores are compared.

The scope of this paper excludes the more general problems of selection of predictors, relative weighting of predictors, and predictive validity of the WMS as now defined. In addition, the problem of accounting for differences in high schools is excluded.

## COMPUTATION OF CLASS STANDING

### PERCENTILE RANK FORMULA

Because class ranks are not comparable across classes of different sizes, they must be transformed to a scale which permits fair comparisons among students: the percentile rank, which indicates the proportion of other students in the class that a particular student exceeds. This may be computed for class rank R and class size N as $(N - R + .5)/N$ or $(2N - 2R + 1)/2N$. The complementary proportion of other students in the class that exceed a particular student may be computed as $(R - .5)/N$ or $(2R - 1)/2N$. A proportion may be converted to percentile rank by multiplying by 100.

The psychometric rationale underlying development of these formulas is documented by Ghiselli (1964), and Guilford (1954). The formulas are modifications of expressions for class rank relative to class size $(N - R)/N$ and $R/N$. Because class rank is a discrete manifestation of a latent trait continuum, each individual actually occupies a range of percentile scores. The expression for class rank relative to class size is adjusted by adding or subtracting .5 to place the individual at the midpoint of the percentile interval.

## DISTRIBUTIONAL ASSUMPTION

An important feature of percentile ranks is that the percentile units represent equal numbers of students rather than equal amounts of the latent trait measured by class rank. By definition, the frequency distribution of percentile ranks is uniform or rectangular in shape. If the distribution of the latent trait is also rectangular, the percentile rank provides an adequate scale for comparison of students in terms of relative amounts of the latent trait. If the distribution is nonrectangular, then percentile rank may grossly distort individual differences among students.

For example, if the latent trait distribution is unimodal and symmetrical, percentile rank would exaggerate individual differences among students near the center of the distribution compared to those near the extremes of the distribution. As a result, use of percentile ranks may place too much emphasis on individual differences among average students and too little emphasis on individual differences among exceptional students for whom class rank may discriminate more finely. For further discussion and an illustration of this problem, consult Guilford (1954, 1956) and Ghiselli (1964).

To compare class standing scores in terms of relative amounts of the latent trait presumed measured by class rank, a theoretical model of the underlying trait distribution must be adopted. A wide variety of distributions is available as potential candidates for the model. A strong argument may be made for adopting the normal distribution as the model (Ghiselli, 1964). Because the normal distribution is a traditional model in latent trait theory, it is adopted here.

A method of transforming ranks into normalized standard scores with mean $\mu$ and standard deviation $\sigma$ is documented by Ghiselli (1964). To facilitate use of this method, Table 1 presents selected percentile rank ranges with equivalent T-scores (i.e., normalized standard scores with $\mu = 50$ and $\sigma = 10$) adapted from Ghiselli (1964). To permit a user to bypass percentile rank computation, Figure 1 gives a plot of several selected T-score contours as a function of class rank and class size. A listing of the FORTRAN computer program used to generate the contour points is provided in the Appendix.

Table 1

**Ranges, Standard Scores, and Total Points for High School Class Standing Formula**

| Range | Score | Weighted Score | Range | Score | Weighted Score |
|---|---|---|---|---|---|
| .0000 - .0015 | 80 | 240 | .4816 - .5185 | 50 | 150 |
| .0016 - .0025 | 79 | 237 | .5186 - .5585 | 49 | 147 |
| .0026 - .0035 | 78 | 234 | .5586 - .5975 | 48 | 144 |
| .0036 - .0045 | 77 | 231 | .5976 - .6345 | 47 | 141 |
| .0046 - .0055 | 76 | 228 | .6346 - .6725 | 46 | 138 |
| .0056 - .0075 | 75 | 225 | .6726 - .7075 | 45 | 135 |
| .0076 - .0095 | 74 | 222 | .7076 - .7405 | 44 | 132 |
| .0096 - .0125 | 73 | 219 | .7406 - .7715 | 43 | 129 |
| .0126 - .0155 | 72 | 216 | .7716 - .8015 | 42 | 126 |
| .0156 - .0205 | 71 | 213 | .8016 - .8275 | 41 | 123 |
| .0206 - .0255 | 70 | 210 | .8276 - .8515 | 40 | 120 |
| .0256 - .0325 | 69 | 207 | .8516 - .8735 | 39 | 117 |
| .0326 - .0405 | 68 | 204 | .8736 - .8935 | 38 | 114 |
| .0406 - .0495 | 67 | 201 | .8936 - .9105 | 37 | 111 |
| .0496 - .0615 | 66 | 198 | .9106 - .9255 | 36 | 108 |
| .0616 - .0745 | 65 | 195 | .9256 - .9385 | 35 | 105 |
| .0746 - .0895 | 64 | 192 | .9386 - .9505 | 34 | 102 |
| .0896 - .1065 | 63 | 189 | .9506 - .9595 | 33 | 99 |
| .1066 - .1265 | 62 | 186 | .9596 - .9675 | 32 | 96 |
| .1266 - .1485 | 61 | 183 | .9676 - .9745 | 31 | 93 |
| .1486 - .1725 | 60 | 180 | .9746 - .9795 | 30 | 90 |
| .1726 - .1985 | 59 | 177 | .9796 - .9845 | 29 | 87 |
| .1986 - .2285 | 58 | 174 | .9846 - .9875 | 28 | 84 |
| .2286 - .2595 | 57 | 171 | .9876 - .9905 | 27 | 81 |
| .2596 - .2925 | 56 | 168 | .9906 - .9925 | 26 | 78 |
| .2926 - .3275 | 55 | 165 | .9926 - .9945 | 25 | 75 |
| .3276 - .3655 | 54 | 162 | .9946 - .9955 | 24 | 72 |
| .3656 - .4025 | 53 | 159 | .9956 - .9965 | 23 | 69 |
| .4026 - .4415 | 52 | 156 | .9966 - .9975 | 22 | 66 |
| .4416 - .4815 | 51 | 153 | .9976 - .9985 | 21 | 63 |
| | | | .9986 -1.0000 | 20 | 60 |

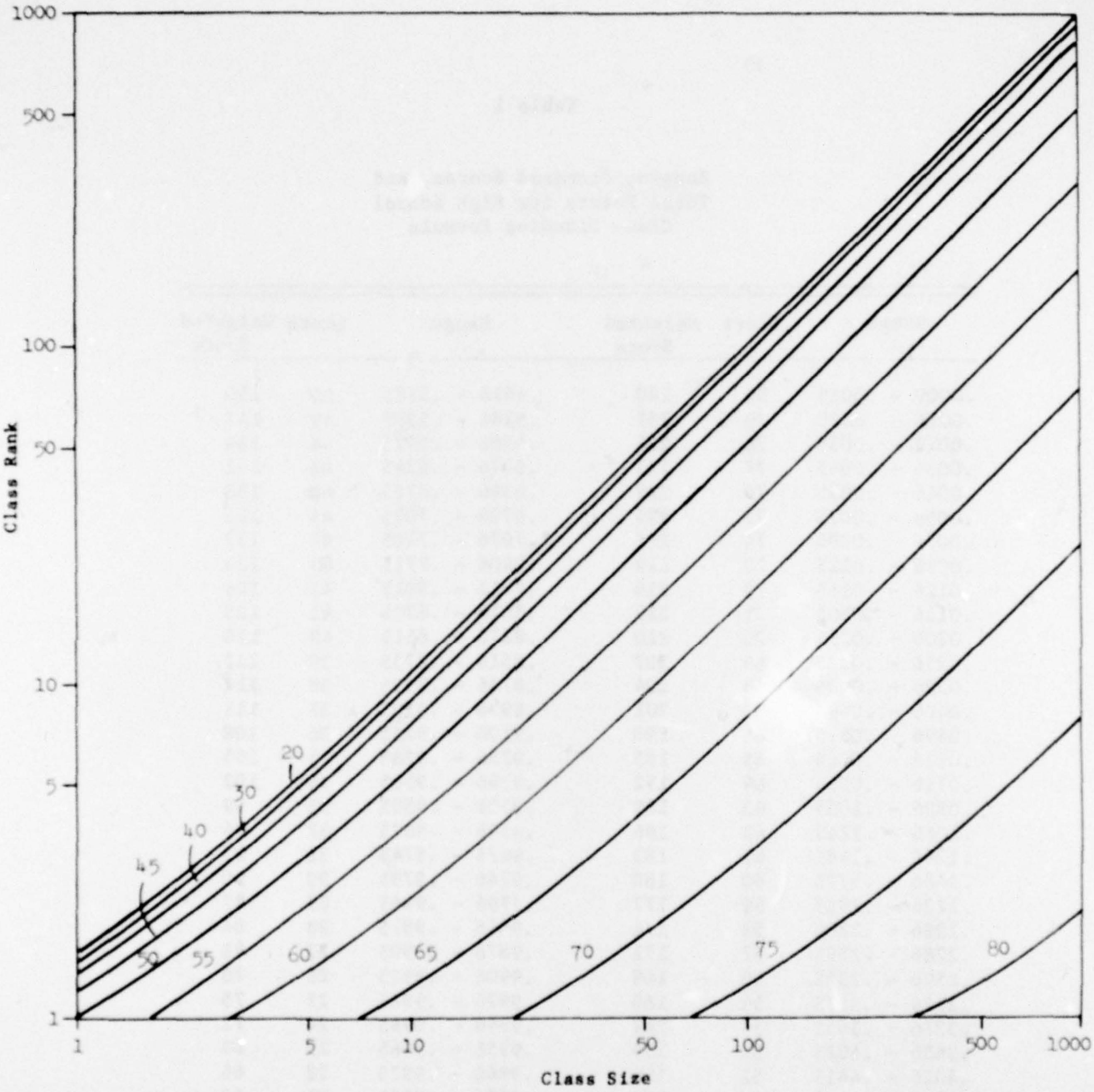


**Figure 1. Normalized Standard Score Contours as a Function of Class Rank and Class Size ($\mu = 50$, $\sigma = 10$).**

Figure 1 provides good discrimination among students in the top half of the class. If good discrimination among students in the bottom half of the class is desired, the following procedure may be used for class rank R and class size N:

1. Enter Figure 1 with class size N and class rank $R^* = N - R + 1$.

2. Read the resulting $T^*$ value from Figure 1 and compute $T = 100 - T^*$ to obtain the correct T value for class rank R and class size N.

## MISSING DATA ESTIMATION

Occasionally a scholarship applicant's high school transcript does not provide class rank and class size information. To compute the applicant's WMS, the missing class standing data must be estimated. The estimation procedure should be as fair as possible to both the applicant and other competitors. Two special cases may be distinguished.

Case I. Partial Data Available. Although the transcript does not provide class rank with class size, it may provide a quartile, decile, percentile (i.e., centile), or similar grouping. Thus, partial information may be available for generating an estimate of actual class standing. If the transcript provides a perce tile rank for an applicant, Table 1 may be used to obtain the class anding score directly. If a quartile, decile, or similar grouping is provided, it is necessary to compute the midpoint (i.e., median percentile) of the grouping prior to entering the table.

Case II. No Data Available. In this case the transcript provides neither class rank with class size nor a quartile, decile, percentile, or similar grouping. No information is available as a basis for estimating actual class standing. In this instance, two alternative approaches are available for generating an estimate of the class standing score. Several methods within each approach are compared below.

The Whole Man Score is redefined here as $WMS = aA + bB + cC + dD$ where a, b, c, and d are weights assigned to predictors A, B, C and D, respectively. By convention. the SAT/ACT predictor is labeled A, the extracurricular/athletic/leadership predictor B, the interview predictor C, and the class standing predictor D; the arithmetic means of WMS, A, B, C, and D are denoted by $\overline{WMS}$, $\bar{A}$, $\bar{B}$, $\bar{C}$, and $\bar{D}$, respectively. The class standing estimate is denoted by $\hat{D}$.

In the first approach, a constant value is computed and used as the estimated class standing for all applicants with missing class standing. A variety of methods are available for computing the constant value:

(1) Set $\hat{D}$ = central tendency measure of latent trait distribution;

(2) Set $\hat{D} = \bar{D}$;

(3) Set $\hat{D} = (a\bar{A} + b\bar{B} + c\bar{C})/(a + b + c)$;

(4) Set $\hat{D} = \overline{WMS} = (a\bar{A} + b\bar{B} + c\bar{C} + d\bar{D})/(a + b + c + d)$;

where the means are computed using data for all applicants with available D. These means may be computed on the basis of the current year's applicants or the previous year's applicants, depending on operational expedience.

In Method 1, knowledge of the form of the latent trait distribution is required. The missing class standing estimate is taken to be the value of some measure of central tendency (mode, median, or mean) of the population distribution of the latent trait. If scholarship applicants represent a random sample from a normal distribution, then mode = median = mean = $\mu$ of the normalized standard score distribution. Because scholarship applicants come predominantly from the top half of the distribution, $\mu \leq$ mode < median < mean of the truncated normalized standard score distribution (Guilford, 1956). Thus, the choice of the measure of central tendency becomes important in practice.

The mode may be rejected immediately. If the normalized standard score distribution is truncated below $\mu$, then mode = $\mu$; if the distribution is truncated above $\mu$, then the mode is the lowest possible score in the truncated distribution. Thus, the mode provides an extremely conservative estimate of true class standing. A better estimate would be offered by a measure of central tendency which places an individual nearer the center of the truncated trait distribution.

One possible estimate is the median normalized standard score of the truncated latent trait distribution. Because the median score by definition corresponds to the 50th percentile of the frequency distribution associated with the truncated latent trait distribution, use of the median involves a distortion problem. The average deviation of true scores from the median is less for students below the median than for students above the median of the truncated latent trait distribution. In other words, the median is a more accurate estimator of true class standing for students in the lower half of the truncated latent trait distribution than for students in the upper half. A better estimate would be offered by a measure of central tendency which is equally accurate for students above and below the estimate.

Such an estimate is the mean normalized standard score of the truncated latent trait distribution. By definition, the mean is the point from which the sum or average of deviations of true scores is the same above and below the point. Moreover, the sum of squared deviations about the mean is smaller than the sum of squared deviations about any other point in the distribution. Thus, the mean is the value which best approximates the true scores of the truncated latent trait distribution as a whole.

To compute the mean normalized standard score of the truncated latent trait distribution, the point of truncation must be specified precisely. In practice, the truncation point may vary from year to year and may not be determined with complete accuracy. Also, the actual latent trait distribution of the applicants may not conform exactly to the truncated normal distribution. Although the actual distribution may conform well down to a certain truncation point, a few students in the truncated portion of the curve may apply. This would result in a small tail on the lower side of the distribution. If some of these applicants have missing class standing scores, using the mean of the truncated latent trait distribution as the estimate of missing class standing would result in a bias in favor of these inferior students.

One way to circumvent this problem is to use the actual mean of normalized standard scores computed on the basis of students with available class standing as in Method 2. The problems of a fluctuating truncation point and a partial tail of applicants below the truncation point may be minimized by recomputing the missing class standing estimate each year. Whatever the form of the latent trait distribution, the mean is the value which best approximates the true scores.

Methods 3 and 4 provide no apparent advantage over Method 2 and have several additional computational disadvantages. Method 3 redistributes the weight d assigned to the missing class standing over the means $\bar{A}$, $\bar{B}$, and $\bar{C}$ of the nonmissing variables in proportion to weights a, b, and c, respectively. This results in $\hat{D}$ being an estimate of WMS minus the D component. In Method 4, the estimate is improved by including $\bar{D}$ in the computation of $\hat{D}$. The approximation is not exact because $\bar{A}$, $\bar{B}$, $\bar{C}$, and $\bar{D}$ are based only on applicants with available D. Both methods suffer the disadvantage that characterized the median: the resulting class standing estimate may be a better estimator for one group of students than for another.

An important feature of the first approach is that all applicants with missing class standing are placed on an equal footing in the competition, inasmuch as all variation in WMS for these applicants is due to the remaining predictor variables. An applicant with missing class standing, however, may be at an advantage or disadvantage with respect to applicants with available class standing depending on whether (a) the constant chosen as the estimate is greater or less than the applicant's true class standing, and (b) the contribution of class standing is large or small relative to the contributions of the other predictors to the overall variation of WMS. Differential effects of condition (a) are minimized in Method 2. A more refined approach is required to minimize differential effects of condition (b).

In the second approach, a different value of estimated class standing is computed for each applicant with a missing class standing. The objective in computing the estimate is to use information about available scores of an applicant to reflect the true class standing. Two potentially useful methods of accomplishing this objective are:

(5) Set $\hat{D} = (aA + bB + cC)/(a + b + c)$;

(6) Set $\hat{D} = (a'A + b'B + c'C)/(a' + b' + c')$;

where a', b', and c' are weights derived via regression of D onto A, B, and C using data for all applicants with available D. The regression weights may be updated annually on the basis of the current year's applicants or the previous year's applicants, depending on operational expedience.

For these methods to provide valid estimates of class standing, class standing must be substantially correlated with the linear combination of other predictors so that some true score variance is recovered. The effect of Method 5 is to redistribute the weight d assigned to D over the available A, B, and C scores in proportion to weights a, b, and c. Although Method 5 is simple, straightforward, and operationally expedient, a possibility exists that this linear combination of other predictors may not account for a substantial amount of true score variance. Method 6 provides the advantage of empirically determined optimal weights which recover a maximum amount of true score variance. Thus, for statistical reasons, Method 6 may be preferred to Method 5.

Our only remaining task is to compare the preferred Method 2 of the first approach with the preferred Method 6 of the second approach. In terms of operational expedience, Method 2 is superior to Method 6, which requires a separate set of standard score conversion tables for A, B, and C to estimate points assigned to D. The increased administrative load may be worthwhile, however, if Method 6 can recover a substantial proportion of true score variance and if a substantial number of applicants have missing class standings. In evaluating this expediency-accuracy tradeoff, data are required to determine empirically the proportion of applicants with missing class standing and the amount of true score variance that can be recovered by Method 6.

To evaluate the statistical utility of Method 6, data were obtained on all male ROTC scholarship applicants for school year 1974. Only 9 out of 2186 applicants did not have high school class standing data available. The intercorrelation matrix for (A) SAT/ACT score, (B) extracurricular/athletic/leadership score, (C) interview score, and (D) class standing score is presented in Table 2 for the remaining 2177 applicants. The multiple regression of D onto A, B, and C yielded a vector of weights $(a', b', c') = (.0075, -.1967, -.1050)$ with associated $R^2 = .0566$. Because only a very small fraction of applicants had missing class standing, and because Method 6 accounted for only 6% of the true class standing variance, it seems reasonable to adopt Method 2 rather than Method 6.

Over all, there is a minimal degree of association among the WMS components, suggesting that each component potentially contributes unique information to the selection procedure for ROTC scholarship applicants. The size of the standard deviations in Table 2 suggests that the lack of association among components is not caused by restriction of range. Regardless of whether the information provided by the four components is indeed unique, the validity of that information for selection of ROTC scholarship recipients is not addressed by the present data, and the issue remains to be investigated.

Table 2

Standard Score Means, Standard Deviations, and Intercorrelations for Components of the Whole Man Score (N = 2177)

| | | | Intercorrelations | | | |
|---|---|---|---|---|---|---|
| | Mean | SD | A | B | C | D |
| A | 178.8 | 18.3 | 1.00 | -.30 | -.09 | .08 |
| B | 169.6 | 22.0 | | 1.00 | .14 | -.21 |
| C | 73.4 | 7.9 | | | 1.00 | -.13 |
| D | 198.9 | 15.9 | | | | 1.00 |

NOTE. A, SAT/ACT score; B, extracurricular/athletic/leadership score; C, interview score; D, class standing score.

SUMMARY

This research memorandum is an outgrowth of technical advisory service performed for the U.S. Army Training and Doctrine Command (TRADOC), Fort Monroe, Virginia in FY 1975. The paper documents the psychometric rationale underlying the method of computing the high school class standing portion of the Whole Man Score (WMS) criterion for selection of Army ROTC scholarship recipients. The program as a whole is designed to provide a comprehensive evaluation of the Army ROTC scholarship recipient selection program.

High school class standing can be adjusted for class size by converting class rank to percentile rank. Percentile rank is then transformed to a corresponding normalized standard score under the traditional assumption of an underlying normal distribution of the latent trait measured by class standing. Several methods of estimating missing class standing scores are compared. A table of normalized standard score equivalents for various percentile rank ranges facilitates use of the method. A plot of selected normalized standard score contours as a function of class rank and class size permits the user to bypass computation of percentile rank. The Appendix lists the FORTRAN computer program used to generate the contour points.

REFERENCES


Ghiselli, Edwin E. Theory of Psychological Measurement. New York: McGraw-Hill, 1964.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1954.

Guilford, J. P. Fundamental Statistics in Psychology and Education. New York: McGraw-Hill, 1956.

U.S. Army Training and Doctrine Command. *TRADOC Circular* 145-5, Reserve Officer's Training Corps 1975-76 Army ROTC Scholarship Administrative Instructions. 15 August 1975.

APPENDIX

FORTRAN PROGRAM USED TO GENERATE NORMALIZED STANDARD SCORE CONTOURS

```
      PROGRAM MAIN
      REAL LIMITS (37)
      DATA ((LIMITS(I), I=1,37)=1.0000, .9945, .9795, .9385, .8515,
     *   .7075, .5185, .4815, .4415, .4025, .3655, .3275, .2925, .2595,
     *   .2285, .1985, .1725, .1485, .1265, .1065, .0895, .0745, .0615,
     *   .0495, .0405, .0325, .0255, .0205, .0155, .0125, .0095, .0075,
     *   .0055, .0045, .0035, .0025, .0015)
1001  FORMAT (1H1/22H CONTOUR LINE FOR T = ,I2, 17H (POINTS = T*3 = ,I3,
     * 9H) LIMIT = ,F6.4//25H (CLASS RANK, CLASS SIZE))
1002  FORMAT (/1H ,3X,F7.2,7X,I4)
      IT=15
      INCR=5
      DO 200 I=1,37
      C=LIMITS (I)
      IT=IT+INCR
      IF(I.EQ. 7)INCR=1
      IP=IT*3
      WRITE (1,1001)IT,IP,C
      DO 100 J=1,3
      M=10**(J-1)
      DO 100 K=1,9
      N=K*M
      R=C*N+0.5
      WRITE (1,1002) R,N
 100  CONTINUE
      N=1000
      R=C*N+0.5
      WRITE (1,1002) R,N
 200  CONTINUE
      STOP
      END
```

PRECEDING PAGE BLANK-NOT FILMED